The role of prominence scales for the disambiguation of grammatical functions in Russian

Lobanova, Anna

Published in:
Russian Linguistics

DOI:
10.1007/s11185-010-9066-3

Russ Linguist (2011) 35: 125–142
DOI 10.1007/s11185-010-9066-3

The role of prominence scales for the disambiguation of grammatical functions in Russian

Роль шкал проминантности в дифференциации грамматических функций в русском языке

Anna Lobanova

Published online: 5 January 2011

Abstract Recent studies on Norwegian, German, and English show that the ordering of constituents in transitive sentences depends on their animacy, definiteness, pronominalization and length. It has further been suggested that these properties can be used to predict grammatical functions of NPs. We examine whether these properties play the same role in Russian, a language with a rather free word order and a rich morphologically-marked case system.

In a corpus-based study, we analyzed 300 SVO and 300 OVS sentences taken from a novel and a newspaper. The results suggest that animacy and pronominalization can be used to predict the position of constituents, but not their grammatical functions. When the preverbal position coincided with the subject position (SVO), the probability of animate NP to be the subject was the same as its probability to be initialized. Pronominalization was a reliable indicator of subjecthood in SVO sentences but a strong predictor of objecthood in OVS sentences. Thus, when case-marking distinguishes between grammatical functions, word order primarily indicates information structure allowing marked constituents in a marked OVS order. This is not taken into account by approaches that use such properties for the disambiguation of grammatical functions.

Аннотация Недавние исследования в норвежском, немецком и английском языках показали, что порядок слов в переходных предложениях зависит от одушевленности, определенности и длины составляющих их фразовых групп. Более того, было предложено, что эти признаки можно использовать для определения грамматических функций именных групп. Данная статья посвящена анализу этих признаков и их влиянию на порядок слов в русском языке, характеризующимся относительно свободным порядком слов и богатой морфологической системой падежей.

I am deeply grateful to Henk Zeevat for bringing up this topic to my attention and patiently waiting for the results. I am also thankful to Gerlof Bouma, Peter de Swart and Jennifer Spenader for reading earlier versions of the paper and providing extensive comments and to Maria Filiouchkina-Krave and Elena Tribushinina for giving useful comments on a short notice. Finally, I am sincerely thankful to the anonymous reviewers.

A. Lobanova (✉)
Artificial Intelligence Department, University of Groningen, Groningen, The Netherlands
e-mail: a.lobanova@ai.rug.nl

В корпусном исследовании были рассмотрены 300 SVO и 300 OVS предложений из нарратива и газеты. Результаты исследования продемонстрировали, что в русском языке одушевленность и прономинализация указывают позицию именных групп, но не определяют их грамматические функции. Так, только когда предглагольная позиция совпадает с позицией подлежащего (SVO предложения), одушевленность может быть использована для определения грамматических функций именной группы. Прономинализация оказалась существенным индикатором подлежащего в SVO предложениях, но еще более значимым индикатором дополнения в OVS предложениях. Данные результаты подчеркивают, что в языках, в которых падежи различают подлежащее и дополнение, порядок слов отражает поток информации, в результате чего фразовые группы с нетипичными признаками разрешены в маркированном порядке слов (OVS). Эта вариация не учтена существующими теориями, рассматривающими одушевленность и определенность как основные признаки различия грамматических функций именных групп.

1 Introduction

Languages differ as to the mechanisms they employ for the disambiguation of grammatical functions. In languages with no or partial case-marking, word order is often the main indicator of grammatical functions. For example, in English transitive sentences, subjects always occupy the pre-verbal position while objects follow the verb. As a result, word order in such languages is rather strict. In languages like Russian, on the other hand, subjects are generally distinguished from objects by means of obligatory case marking. Word order can therefore be used for other purposes, e.g. to indicate the information structure of a sentence. Because of that, Russian word order is rather flexible.

Recent studies have argued that animacy and definiteness also play a role in the disambiguation of grammatical functions since subjects tend to be animate and definite whereas objects tend to be inanimate and indefinite (see Dahl 2000 on Swedish; Weber and Müller 2004 and Heylen 2005 on German; Bouma 2008 on Dutch). In particular, it has been shown that non-typical combinations like animate human pronoun objects are very infrequent in English and Swedish corpora (Zeevat and Jäger 2002). Consequently, knowing that an NP is human can be used to predict that it is likely to be the subject. Also, the distribution of properties over constituents seems to play an important role for disambiguation. For example, Øvrelid (2004) has shown based on a corpus of Norwegian, an SVO language, that when objects were more animate than subjects, they never preceded subjects. In other words, because animacy is not a typical property of objects, animate objects avoided marked OVS position.

The findings described above are important not only for the theory of language but also for computational applications such as automatic parsing. Since animacy and definiteness are universal properties of subjects and objects, similar findings should be expected across different languages. So far, however, the focus of previous work was on languages with a relatively fixed word order. The main goal of this study is to examine whether animacy and definiteness play a similar role in Russian, a language with a free word order. Since case marking in Russian takes care of disambiguation between grammatical functions, and the ordering of constituents directly reflects its information structure, this language offers an opportunity to investigate whether subjects are more animate and definite regardless of their position in a sentence or whether fronted constituents are more animate and definite regardless of their grammatical functions. In other words, do animacy and definiteness

of constituents interplay with grammatical functions (as has been previously assumed) or rather with the information structure? We will examine the following subparts of this question:

Q1: In Russian transitive sentences, are subjects more animate and definite than objects?
Q2: In Russian transitive sentences, do non-typical constituents (animate, pronominalized objects) avoid marked (OVS) order?
Q3: In Russian transitive sentences, is it possible to predict grammatical functions of NPs given their properties?

To answer these questions, we analyzed the properties of subjects and objects in a total of 600 SVO and OVS Russian transitive sentences. We examined the overall animacy and definiteness of subjects and objects as well as their animacy and definiteness in respect to each other. Our main finding is that unlike in English and Swedish, non-typical constituents in Russian tend to appear in marked orderings. Animacy and definiteness seem to primarily interplay with the information structure rather than the disambiguation of constituents and, as a consequence, these properties cannot be used to predict grammatical functions of NPs in Russian transitive sentences.

The rest of the paper is organized as follows. In Sect. 2 we introduce the default word order in Russian; we explain typical properties of subjects and objects in regard to their prominence on animacy and definiteness scales and we discuss how these properties have been used to predict the grammatical functions of NPs in languages with a strict word order. The current study is presented in Sect. 3. The results are presented in Sect. 4. In Sect. 5, we discuss the findings; conclusions are summarized in Sect. 6.

2 The nature of word order variation

2.1 Default word order

Languages differ in the amount of freedom in word order variation they allow. In English transitive sentences, the word order is fixed, and the subject always precedes the verb while the object tends to follow it. In German, word order is partially free allowing some variation (SVO, OVS). Free word order languages like Russian allow all possible variations.

It is generally acknowledged that SVO is the default word order in Russian (Kovtunova 1976). Sentence (1), where the subject precedes the verb and the object follows it, is an answer to a general question 'What is happening?' when all information is new:

(1) SVO
Мария убирает комнату.
Marija.Nom is cleaning room.Acc
'Maria is cleaning a room.'

However, this word order is not fixed and the other five possible variations (i.e. OVS, SOV, OSV, VSO and VOS) are allowed as well. For example, in (2) the object is fronted, but the interpretation is not ambiguous because of the case-marking on both constituents.

(2) OVS
Комнату убирает Мария.
room.Acc is cleaning Marija.Nom
'It is Maria who is cleaning the room.'

What then is the difference between these two sentences? The sentences are interpreted differently due to the information structure they convey. In (1), when *Мария* 'Maria' is given information and *убирает комнату* 'is cleaning a room' is new information, this sentence is an answer to the question 'What is Maria doing?'. In (2), the subject is focused and it conveys new information while the object is given. It is an answer to 'Who is cleaning the room?'. The definiteness of the object is thus another difference between (1) and (2). Since there are no articles in Russian, word order in these examples marks definiteness.[1]

What happens when morphology does not distinguish between grammatical functions? Such sentences are interpreted as SVO. Consider the well-known example in (3):

(3) SVO/OVS (ambiguous)

Мать	любит	дочь.
mother.Nom/Acc	loves	daughter.Nom/Acc

Literally: 'Mother loves daughter.' (Jakobson 1971[1936], 28)

In (3), the grammatical functions of *мать* 'mother' and *дочь* 'daughter' are ambiguous, as the Nominative and Accusative forms in this case are identical. Nevertheless, any speaker of Russian will interpret this sentence as SVO unless *дочь* is prosodically stressed.[2] This is because in an unmarked transitive sentence in Russian, the initial position is associated with the subject and the post-verbal position with the object. In fact, there is a general tendency cross-linguistically for subjects to precede objects—subjects precede objects in over 90% of the languages (Hawkins 1983). Which factors seem to correlate with word order variation? It has been proposed that animacy, definiteness, referential form and the length of constituents play a role in their ordering. These properties are discussed in the next section.

2.2 Prototypical properties of subjects and objects

Grammatical prominence of subjects and objects is related to their semantic properties. Animacy is a prototypical property of subjects. Dahl (2000) reported that in a corpus of spoken Swedish, 92% of the transitive subjects are animate. According to Jacobsen (1992), in transitive sentences in Japanese, inanimate subjects are not allowed at all. Prototypical subjects are also definite and specific. They are more likely to convey given information and they tend to occupy the initial position in a sentence. Meanwhile, prototypical objects are inanimate, indefinite, non-specific and often introduce new information.

In (1), repeated below as (4), the subject is animate and definite, it is expressed by a proper name. The object is inanimate and indefinite, it is expressed by a noun. Both constituents have prototypical properties and their ordering is unmarked.

[1] Word order is one of the means of indicating definiteness in Russian. To illustrate the interplay between word order and definiteness, consider the examples below where *ваза* 'vase' becomes definite when it is fronted as in (ii):

(i) На столе стоит ваза.
on table stands vase.Indef
'There is a vase on the table.'

(ii) Ваза стоит на столе.
vase.Def stands on table
'The vase is on the table.'

[2] When *дочь* is prosodically stressed, (3) is interpreted as an OVS sentence. Note further that King (1995) argues that contexts also make an OVS reading of this sentence possible (suggesting that word order in Russian does not indicate grammatical functions).

(4) [What is happening?]
Мария убирает комнату.
Marija.Nom is cleaning room.Acc
'Maria is cleaning a room.'

The properties discussed above can be expressed by means of prominence scales (Aissen 2003). The dimension of animacy is represented as a scale (5), according to which animate nouns are more prominent than inanimate. Definiteness is represented by means of scale (6), according to which pronouns are more prominent than proper names, which are in turn more prominent than definite nouns, and so on.

(5) Animacy Scale: Human > Animate > Inanimate

(6) Definiteness Scale: Pronoun > Proper Name > Definite > Indefinite Specific > Non-specific

The importance of these scales in linguistics has been discussed (among others) in relation to the choice between the Saxon genitive and the *of*-genitive (Rosenbach 2002), between active and passive voice (Rosenbach 2003) and between pronominal and full noun reference (Dahl and Fraurud 1996). Aissen (2003) has convincingly shown that prominence hierarchies play a role in the markedness of overt subject/object case marking. Using 'harmonic alignment'[3] she unified the animacy and definiteness scales with the grammatical functions scale, cf. (7). The obtained scales in (8) denote that human and animate subjects are more prominent than inanimate subjects whereas inanimate objects are more prominent than animate objects. According to (9), subjects expressed by pronouns, proper names and definite NPs are more prominent than subjects expressed by indefinite nouns. In reverse, indefinite objects are more prominent than the ones expressed by definite nouns, proper names and pronouns.

(7) Grammatical Functions Scale: Subject > Object

(8) a. Subject/Human > Subject/Animate > Subject/Inanimate
b. Object/Inanimate > Object/Animate > Object/Human

(9) a. Sbj/Pro > Sbj/PN > Sbj/Def > Sbj/Spec > Sbj/Nspec
b. Obj/Nspec > Obj/Spec > Obj/Def > Obj/PN > Obj/Pro[4]

In summary, animate, definite, pronominal subjects are more typical and therefore less marked than the ones that are inanimate and indefinite. Inanimate indefinite objects expressed by full NPs are more typical and therefore less marked than pronominalized animate objects (Givón 1983, 2001). According to this approach, what is marked for subjects

[3] 'Harmonic alignment' was introduced by Prince and Smolensky (1993) as part of their account of the relation between sonority and syllable structure. The underlying idea is to correlate pairs of scales, aligning each element on one scale with each element on the other. If there is a dimension D1 with a scale X > Y and another dimension D2 with scale a > b > … > z, the harmonic alignment of D1 and D2 is the pair of harmony scales (where '»' means 'less marked than'):

X/a»X/b» … »X/z

Y/z» … »Y/b»Y/a

[4] Sbj = subject, Obj = object, Pro = pronoun, PN = proper name, Def = definite, Spec = indefinite specific, Nspec = non-specific and F = focus.

is unmarked for objects and vice versa what is marked for objects is unmarked for subjects. This has been referred to as 'markedness reversal' (Battistella 1990; Aissen 2003; Comrie 1989).

If scales play a role in the prominence of subjects and objects, it should be reflected in the data. Zeevat and Jäger (2002) used 250,000 noun phrases from the Wall Street Journal Corpus of English (as a part of the Penn Treebank corpus; cf. also Marcus et al. 1993) and a corpus of taped and transcribed everyday conversations in Swedish to look at the frequencies of non-typical combinations like human pronominalized objects. Although the two corpora were different (there were fewer pronouns and animate nouns in the newspaper corpus), they found that disharmonic combinations were dispreferred in both languages. For example, in English, the probability of an object to be human (42%) was lower than the probability of an object to be expressed by a noun (75%); the probability of a human NP to be an object (10%) was lower than the probability of a noun phrase to be human (13%). Based on these results, Zeevat and Jäger (2002) argued that given the animacy and definiteness of NPs in a transitive sentence, it is possible to predict their grammatical functions. For the corpus of English, the probability of an NP to be the object of a sentence was 75% and the probability of a pronoun to be the subject was 88%. The probability of an inanimate noun to be an object was 90%. For Swedish, the probability of a human NP or an ego pronoun ('I', 'you', 'we') to be the subject of a transitive sentence was 97%, while the probability of a definite noun or a pronoun to be the object was 15% and 17% respectively (which was also lower than the probability of a noun to be the object). These results provide evidence based on corpus frequencies, supporting their claim that animacy and pronominalization are reliable predictors of subjecthood while indefiniteness is a good predictor of objecthood.

In typological linguistics, prominence scales have also been extensively used to explain variations and restrictions on the ordering of subjects and objects in relation to each other (e.g. Siewierska 1988). For example, Morimoto (2001) suggests that in Kinyarwanda, a Bantu language, object fronting is only possible when the subject is more animate—prominent—than the object. Similarly, the corpus data from Øvrelid (2004) suggest that in Norwegian transitive sentences, objects that are more animate than subjects avoid marked (OVS) word order. Out of the 1000 sample sentences, 9.7% had an OVS word order and in none of those sentences objects were higher in animacy than subjects.

Corpus analysis provides evidence based on language usage. However, so far, corpus-based studies that looked at the role of prominence scales in disambiguation of grammatical functions have only been done on languages with relatively strict word orders. Thus, the predictions are based on the assumption that there is a direct relation between grammatical functions and prominence. If this generalization is valid and prominence is the key factor in determining grammatical functions, it should hold in all word order variations. For Russian that would mean that subjects will be more animate and definite than objects, and objects will be more inanimate and indefinite than subjects in both SVO and OVS word order variations. This is an interesting prediction given that the word order in Russian has different functions. Analyzing a corpus of sentences from a free word order language like Russian seems like a necessary next step.

3 Current study

The current work consisted of two parts. First, we determined relative frequency distributions of six word order variations in a sample of the first 300 transitive sentences

extracted from a novel. In the main study, we looked at animacy, referential form and length of subjects and objects in two sets of sentences with the most frequent word order variations—SVO and OVS. To examine whether the results are genre-dependent, these sentences represented two genres—novel and newspaper. Sentences from the novel (Evdokimov 1959) contained a lot of pronouns and proper names as the main characters were human. On the other hand, most of the newspaper articles were devoted to news about world events (meetings, wars, storms, etc.) so we expected to find a difference in the total number of animate NPs in the sentences taken from the novel and from the newspaper. No differences were expected in respect to the ordering of the prominent elements in sentences.

The set of sentences from the novel contained 150 SVO sentences (taken from the total 265 SVO sentences from the original 300) and 150 OVS sentences (17 were taken from the original 300 sentences and 133 sentences were added extra). The newspaper set consisted of 150 SVO and 150 OVS sentences taken from *Аргументы и Факты*.[5]

By comparing the properties of subjects and objects within and between SVO and OVS sentences, we examined how prominence scales play a role in determining the position of subjects and objects in Russian transitive sentences. Further, following Zeevat and Jäger (2002), we looked at whether similarly to English and Swedish the properties of NPs could be reliably used to predict their grammatical functions.

3.1 Word orders distribution

In the preliminary study we looked only at sentences from the novel. They were written in a simple, clear style containing a lot of personal pronouns and proper names. The first 300 transitive sentences were classified as SVO, OVS, OSV, SOV, VSO or VOS. Out of them, 265 sentences (88%) were SVO, 17 (6%) OVS, 11 (4%) OSV, 5 (1.5%) SOV and 2 (0.5%) VOS. These frequency distributions are similar to the results reported in Bivon (1971), who used a corpus of Russian texts from newspapers and novels. In his corpus, 79% of the sentences had an SVO order, 11% OVS, 4% OSV, 2% VOS, 1% VSO/SOV.[6]

Two most frequent word order variations were SVO and OVS. To narrow the scope of the study and due to the problems of collecting enough samples of infrequent word order variations, in the rest of the study we looked at SVO and OVS sentences only.

3.2 Annotation

All sentences were classified as to their animacy (‘animate’ or ‘inanimate’), and definiteness that was subdivided into referential form (‘pronouns’, ‘proper names’ or ‘full noun phrases’) and the length of constituents (‘one word’ or ‘multi-word units’). The results are described in Sect. 4.

3.2.1 Animacy

Despite the recognized importance of the animacy scale, the distinction between what is animate and inanimate is far from clear. It turned out to be a difficult task to properly define the animacy of nouns because the linguistic description of what is animate is not

[5] An electronic archive of the newspaper is available online at http://gazeta.aif.ru/oldsite/.

[6] The remaining 2% included sentences in which one element interrupted another (Bivon 1971, 42).

the same as the biological one. For example, from a biological point of view, 'dog' and 'tree' are both animate. But not linguistically. Another problem is that many nominals are ambiguous—sometimes referring to people and other times to organizations. *Philips*, for example, can refer to a group of people (as in *Philips is on strike*) or to an organization (as in *this year Philips suffered record losses*); *The Netherlands* can refer to a geographical location (as in *your passport must be valid for at least one year after your arrival in the Netherlands*), to the country as a body (as in *The Netherlands has a good infrastructure for an intense amount of rail traffic*) or to its government/organization (as in *The Netherlands and France have held talks about bilateral cooperation*). Inanimate nouns can further be categorized as concrete (e.g. *table*, *tree*, *snow*) or abstract. Abstract nouns refer to events (*the party was fun*) and abstractions (*his future*, *this idea*, *the trip*). In this study, all nouns were first classified as animate or inanimate. However, as will be discussed in Sect. 4.1, this partition was not sufficient and a three-grained division into animate, inanimate abstract and inanimate concrete was used instead. Animate nouns included nominals that referred to humans/animals (*painter*, *cat*, *pigeon*) and groups of people (*family*, *team*); inanimate abstract nouns referred to events, abstractions, locations and countries as a body; and inanimate concrete nouns referred to inanimate concrete objects (*brush*).

3.2.2 Definiteness

We separated definiteness into two dimensions: referential form and length of constituents. Referential form consisted of three categories: pronouns, proper names and nouns. Following the prominence scales, we considered pronouns and proper names to be more definite than nouns. As to the length, constituents were subdivided into one word or multi-word units. One word NPs were considered to be more definite than longer NPs, since the length of constituents is related to the information structure and while given information can be referred to by a pronoun or a one word NP, new information needs introduction and is often described by a phrase that is longer than one word (Givón 1983).

4 Results

4.1 Animacy

Animacy was the first property we examined. Overall there were more animate constituents in the novel than in the newspaper. In the novel, 51% of all NPs in both word order variations were animate. Among those, 60% were subjects and 40% were objects. In the newspaper, 63.3% of all NPs were inanimate. Among those, 40% were subjects and 60% were objects. Out of the 36.7% of animate NPs in the newspaper set, 67% were subjects and 33% were objects. The results on the frequency distribution in respect to animacy for sentences from the novel and the newspaper with regard to the word order variations are given in Fig. 1.

Recall that we expected more animate subjects and inanimate objects in both word order variations. In the sets of SVO sentences we found a significant association between animacy and grammatical functions in both genres: $\chi^2(1) = 59.8$, $p < .001$ for sentences from the novel and $\chi^2(1) = 45.15$, $p < .001$ for sentences from the newspaper. In the novel, 73% of the subjects were animate and 72% of the objects were inanimate; in the newspaper, the distribution of animate and inanimate subjects was even, while 85% of the objects were inanimate. The large number of inanimate subjects in the newspaper set was due to a larger number of inanimate constituents in the newspaper sentences in general.

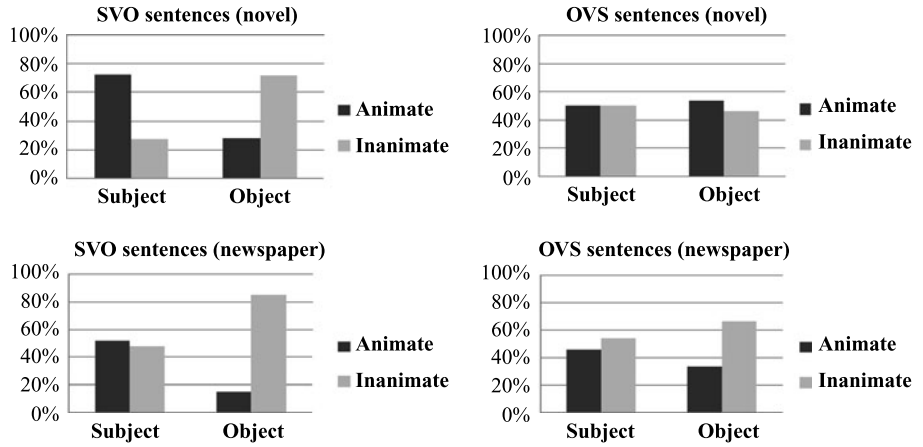

Fig. 1 Frequency distribution of animate/inanimate NPs and their grammatical functions

In the OVS sentences, a significant association between animacy and grammatical functions was found only in the sentences from the newspaper ($\chi^2(1) = 5.028$, $p < .05$). While subjects were evenly animate and inanimate in both OVS sets, there was an increasing number of animate objects, especially in the sentences from the novel, where the number of animate objects was higher than the number of inanimate objects or the number of animate subjects. Compared to SVO sentences, the number of animate objects in the newspaper OVS sentences doubled—from 15% to 33%. This suggests that non-typical animate objects do not avoid marked position.

To examine whether there were differences within inanimate constituents, all inanimate NPs were further subdivided into inanimate abstract and inanimate concrete. Their distribution is illustrated in Fig. 2.

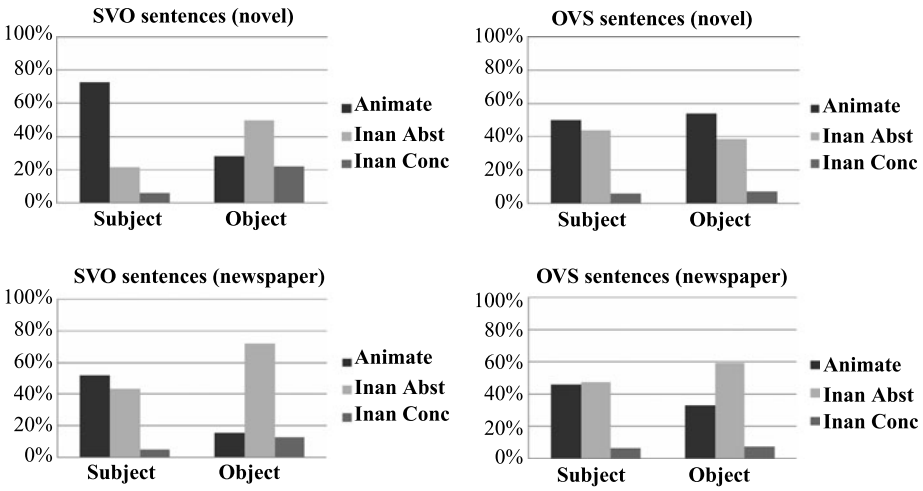

Fig. 2 Frequency distribution of animate, inanimate abstract and inanimate concrete subjects and objects. *Inan Abst*—inanimate abstract, *Inan Conc*—inanimate concrete

There was a big difference between pre-verbal and post-verbal objects regarding the number of inanimate concrete objects. In the novel set, there were three times as many inanimate concrete objects in SVO sentences as in OVS sentences. Similarly, in the newspaper set, there were twice as many inanimate concrete objects in SVO sentences as in OVS sentences. Thus, when pre-verbal objects were inanimate, they were mostly abstract.

But what was the prominence of constituents in respect to each other—were subjects more prominent than objects, or equally prominent? Did word order variation play a role? This distribution is presented in Fig. 3.

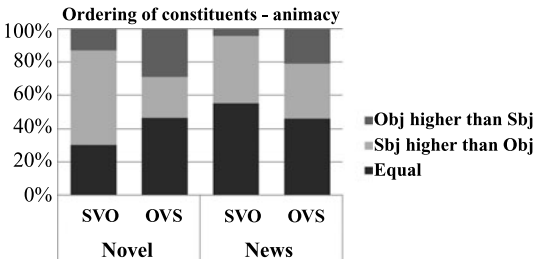

Fig. 3 Proportion of constituents preceding/following each other depending on their animacy

Sbj—subject, *Obj*—object

The distribution of constituents in respect to each other was similar in both genres but differed in respect to word order variation. In SVO sentences, subjects were more animate than objects, or constituents were equally prominent in 87.3% of the sentences from the novel and in 96% of the sentences from the newspaper. In OVS sentences, the number of objects that were more animate (and therefore prominent) increased from 12.7% to 28.7% in the novel set and from 4% to 20.7% in the newspaper set. This again shows that marked objects do not avoid marked OVS position in Russian transitive sentences.

Given that objects were more prominent than subjects in up to 28.7% of OVS sentences, is it still possible to predict grammatical functions of NPs knowing their animacy? As has been discussed earlier, Zeevat and Jäger (2002) showed that in English and Swedish animacy was a reliable predictor of subjecthood because disharmonic combinations were very infrequent and the probability of an object to be animate was very low. The probabilities as to the animacy for the Russian data are given in Table 1.

Table 1 Probabilities of grammatical functions given animacy

	SVO	OVS
Novel	P(Sbj|Anim) = 72%	P(Sbj|Anim) = 49%
	P(Obj|Inan Abs) = 70%	P(Obj|Inan Abs) = 46%
	P(Obj|Inan Conc) = 78%	P(Obj|Inan Conc) = 55%
News	P(Sbj|Anim) = 78%	P(Sbj|Anim) = 57%
	P(Obj|Inan Abs) = 63%	P(Obj|Inan Abs) = 55%
	P(Obj|Inan Conc) = 74%	P(Obj|Inan Conc) = 52%

In both genres, animacy was a good predictor of subjecthood in SVO sentences only. The probability of a noun to be subject, given that it is animate, varied from 72% to 78%.

Inanimacy was a good predictor of objecthood also only in SVO sentences. In OVS sentences, on the other hand, due to the increasing number of non-typical objects, animacy was a reliable indicator neither of subjecthood nor objecthood.

4.2 Referential form

The second property we examined was referential form. We expected, first, more subjects expressed by pronouns and proper names and more objects expressed by full noun phrases; and second, no pronominalized objects in a marked OVS word order. Regardless of genre, the majority of constituents in SVO and OVS sentences was expressed by full noun phrases (60% in the novel and 76% in the newspaper), followed by proper names (22% in the novel and 13% in the newspaper) and pronouns (18% in the novel and 11% in the newspaper). There were more noun phrases and less proper names and pronouns in the newspaper than in the novel. The results on the frequency distribution for sentences from the novel and the newspaper for both word order variations are given in Fig. 4.

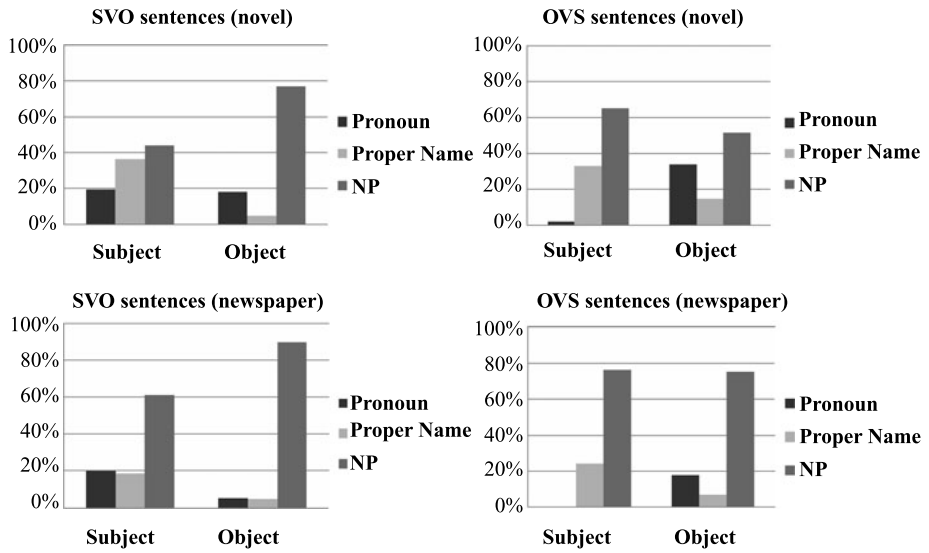

Fig. 4 Frequency distribution of pronouns, proper names and full noun phrases and their grammatical functions

We found a significant association between referential form and grammatical functions in all four sets of sentences. In the sets of SVO sentences $\chi^2(2) = 50.9$, $p < .001$ for sentences from the novel and $\chi^2(2) = 33.4$, $p < .001$ for sentences from the newspaper. In the sets of OVS sentences $\chi^2(2) = 55.4$, $p < .001$ for sentences from the novel and $\chi^2(2) = 41.7$, $p < .001$ for sentences from the newspaper. In the sets of SVO sentences more subjects were expressed by pronouns and proper names and more objects were expressed by full noun phrases. Although the overall frequency of pronouns in SVO and OVS sentences was similar (56 and 54 pronouns in SVO and OVS sentences from the novel and 38 and 27 pronouns in SVO and OVS sentences from the newspaper), pronouns varied drastically as to their grammatical functions. In the SVO sets 50% of the pronouns in the sentences from the novel and 79% of the pronouns in the sentences from the newspaper

were subjects. In the OVS sets, only in 6% of the sentences from the novel and in none of the sentences from the newspapers subjects were pronominalized. The large number of pronominalized objects in OVS sentences shows that non-typical (pronominalized) objects do not avoid fronting. There were also more nominal objects than subjects in SVO sets and more nominal subjects than objects in OVS sets. Proper names, on the other hand, tended to be subjects in both word order variations in both genres.

The distribution of constituents in respect to each other (Fig. 5) was similar in both genres but not in different word orders.

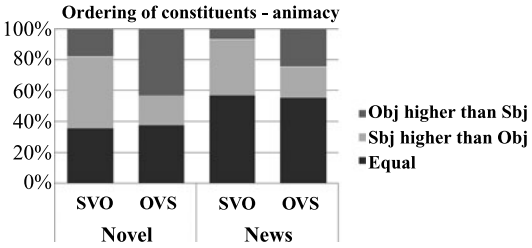

Fig. 5 Proportion of constituents preceding/following each other depending on their referential form

Sbj—subject, *Obj*—object

In SVO sentences, subjects were equally or more prominent on the referential scale in 82% of the sentences from the novel and in 93.4% of the sentences from the newspaper. In OVS sentences, subjects were equally or more prominent than objects in 56.7% of the sentences from the novel and in 75.3% of the sentences from the newspaper. Sentences in which objects were more prominent than subjects were found in all four sets but there were particularly many cases in OVS sentences from the novel—43.3%. In 22% of these sentences subjects were expressed by NPs and objects were expressed by pronouns, in 10.7% of the sentences subjects were expressed by proper names and objects were expressed by pronouns; and in another 10.7% of the sentences subjects were expressed by nouns and objects were expressed by proper names. In the OVS sentences from the newspaper, objects were more prominent than subjects in 24.7% of the sentences. In such sentences subjects were mostly nouns, and objects were either pronominalized or expressed by proper names. Also in SVO sentences from the novel, objects were higher in prominence than subjects in 18% of the cases. Among those sentences subjects were expressed by full NPs and objects were pronominalized (10% of sentences), or expressed by proper names (2.7% of sentences), or subjects were expressed by proper names while objects were pronominalized (5.3% of sentences).

In 19.4% of the OVS sentences from the novel and in 20% of the OVS sentences from the newspaper, subjects were more prominent than objects yet they were post-verbal. In almost all such cases objects were expressed by nominal phrases and subjects were expressed by proper names. This suggests that referential form is not always a defining constraint on the ordering of constituents.

Going back to the frequency distributions of referential forms for subjects and objects in Fig. 4, what do they mean in regard to the possibility to use referential forms of constituents to predict their grammatical functions? The probabilities for the Russian data are presented in Table 2.

The obtained probabilities do not confirm expectations based on the referential scale that subjects are likely to be pronominalized and objects are likely to be expressed by full NPs. Being a proper name was the only reliable predictor of subjecthood regardless of word order variation. The results for pronominalization varied for different word orders and

Table 2 Probabilities of grammatical functions given referential form

	SVO	OVS
Novel	P(Sbj|Pro) = 51%	P(Obj|Pro) = 94%
	P(Sbj|PN) = 89%	P(Sbj|PN) = 69%
	P(Obj|NP) = 63%	P(Obj|NP) = 36%
News	P(Sbj|Pro) = 78%	P(Obj|Pro) = 100%
	P(Sbj|PN) = 80%	P(Sbj|PN) = 79%
	P(Obj|NP) = 59%	P(Obj|NP) = 50%

genres. In SVO sentences, pronominalization was a good indicator of subjecthood only for the set of sentences taken from the newspaper. Contrary to the results in previous studies that showed that pronominalization can be used to predict subjecthood, the probability of a pronoun to be an object in OVS sentences was at least 94%. Also knowing that a constituent is expressed by a full NP would not help to predict its grammatical role as many subjects as well as objects were expressed by nouns.

4.3 Length of NP

The last property we looked at was the length of constituents. In both word order variations subjects were expected to be short more often than objects because they are expected to be more definite. Overall, sentences from the novel contained more short constituents than sentences from the newspaper—52% of subjects and objects in the novel and 34% of subjects and objects in the newspaper were one word long. The distribution frequency of subjects and objects as to their length is given in Fig. 6.

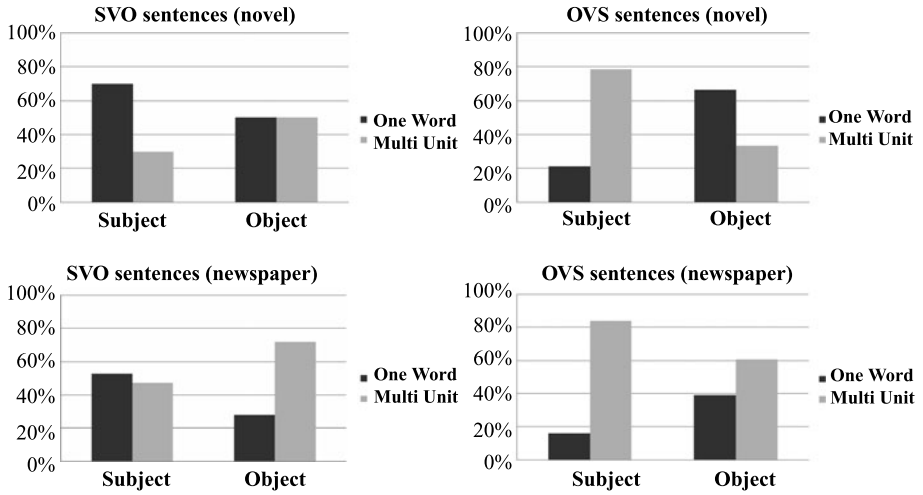

Fig. 6 Frequency distribution of constituents in respect to their length

What was found is that fronted constituents (rather than subjects in particular) tended to be shorter—70% of subjects in the SVO sentences and 67% of objects in OVS sentences were

one word in length. A significant association between length and grammatical functions was found in all sets of sentences ($\chi^2(1) = 12.5$, $p < .001$ for SVO sentences from the novel and $\chi^2(1) = 18.9$, $p < .001$ for SVO sentences from the newspaper; $\chi^2(1) = 20.4$, $p < .001$ for OVS sentences from the novel and $\chi^2(1) = 62.5$, $p < .001$ for OVS sentences from the newspaper). In general, it is plausible to conclude that post-verbal constituents are often lengthier.

5 Discussion

Our main finding is that there is a difference in respect to the role of animacy and definiteness of subjects and objects for the disambiguation of grammatical functions between languages with fixed word order and languages with free word order. While languages with relatively fixed word order, e.g. English, strongly disprefer non-typical constituents to precede typical ones, languages with free word order, in this case Russian, allow variation. Previous studies have shown that animacy is a strong indicator of subjecthood in languages like English and Swedish. Our results suggest that animacy is a less reliable indicator of subjecthood in languages like Russian.

Our second finding is related to the claim that pronominalization can be used to predict subjecthood. Our findings for Russian show that not subjects but fronted constituents in general tend to be pronominalized. The reason for this might lie in the role of word order in Russian as opposed to, e.g., English. Since case-marking indicates grammatical functions in Russian, word order is used instead to reflect the information structure of sentences. Pronouns usually refer to entities that have already been mentioned in discourse. Thus, they are likely to convey given information, and given information tends to precede new information (Clark and Clark 1977; Gundel 1988). As a result, pronominalized constituents in Russian are likely to be fronted regardless of their grammatical function. Although proper names follow pronouns on the definiteness scale, they seem to behave differently in that in the analyzed sample they were likely to be subjects in both SVO and OVS word order variations. This seems to suggest that the 'givenness' of pronouns and proper names differs in that pronouns refer to information that was mentioned earlier in discourse, while proper names can refer to information given in general and not necessarily mentioned before. We will now discuss these findings in more detail.

In respect to animacy, what kind of sentences contained inanimate subjects and animate objects? In our sample most of such sentences contained psych-verbs as their predicates. In some cases such sentences were used in a non-literal meaning. Consider example (10):

(10)

Зимний	холод	угнетал	Левитана.	Снег	тяготил	его.
winter	coldness.Nom	depressed	Levitan.Acc	snow.Nom	oppressed	him.Acc

'The coldness of the winter depressed Levitan. The snow oppressed him.'
(Evdokimov 1959)

The objects of the verbs 'to depress' and 'to oppress' in (10) indicate that 'cold' and 'snow' (the Stimuli) cause an emotional reaction in the subject (or the Experiencer) *Levitan*. Such verbs are called psych-verbs (Levin 1993). Their arguments exhibit a reversed relationship in that the objects are more subject-like while the subjects are more object-like. One way to analyze such predicates is by taking into account Dowty's Selectional Principle. According to this principle, "the argument for which the predicate entails the greatest

number of Proto-Agent properties will be lexicalized as the subject of the predicate; the argument having the greatest number of Proto-Patient entailments will be lexicalized as the direct object" (Dowty 1991, 576). Since the Stimulus in (10) above entails causation (with volition), which is a Proto-Agent property, and the verbs imply a change of state, a Proto-Patient property, the Experiencer becomes a better candidate to be a Proto-Patient.

The tendency to reverse the functions of the arguments of psych-verbs is found across languages, including Norwegian, as is illustrated in (11):

(11) Spørsmålet plager Espen.
question bothers Espen
'The question bothers Espen.' (Øvrelid 2004, 8, modified)

Note that in the Russian sentences in (10) grammatical functions are disambiguated by means of case, so that the subjects are in the Nominative case and the objects are in the Accusative case. In the Norwegian example, on the other hand, the only formal disambiguation between subject and object is done by the word order itself. And because the object in (11) is non-typical and more animate than the subject, reversing the order of constituents leads to degraded grammaticality:

(12) ??Espen plager spørsmålet.
Espen bothers question
??'Espen, the question bothers.' (Øvrelid 2004, 8)

Given the right context, it is possible to interpret (12) as a topicalized version of (11). Still, such an interpretation will be very marked. Pronominalized objects in Norwegian are, however, case-marked. And in such cases, the reverse order of (11) becomes grammatical, see (13):

(13) Meg plager spørsmålet veldig.
me.Acc bothers question very
'Me, the question bothers very much.' (Øvrelid 2004, 8)

Thus, similarly to Russian, when case-marking in Norwegian takes over the role of the disambiguator of grammatical functions, the ordering of constituents in respect to each other becomes more flexible.

The Russian equivalents of (11) and (12), that is, (15a) and (14a), differ in respect to the information structure they convey. In (14) the topic of the sentence is established by a *wh*-question that presupposes that the sentence is about entities that bother Levitan. The assertion that what bothers him is pressing issues, is the focus as it provides new information in relation to the topic. This new information is prosodically stressed. Because the object in this case is part of given information, it can be expressed by definite NPs (including proper names as in (14a)), it can be pronominalized (as in 14b) or it can be omitted (as in 14c):

(14) [What is bothering Levitan?]
a. Левитана беспокоят [$_F$насущные вопросы].
Levitan.Acc bother current questions.Nom
'Pressing issues are bothering Levitan.'
b. Его беспокоят [$_F$насущные вопросы].
him.Acc bother current questions.Nom
'Pressing issues are bothering him.'

c. [$_F$Насущные вопросы].
current questions.Nom
'Pressing issues.'

When the order is reversed, cf. (15), the information structure of the sentence changes so that the new information now is who is bothered by problems. Now that the object contains new information, it can still be expressed by a proper name (15a), but it cannot be expressed by a pronoun (unless the speaker points out to the referent during conversation). Due to their nature, pronouns are likely to contain information that has been mentioned in the discourse itself; therefore they are likely to be the topic of a sentence. Proper names, on the other hand, can refer to entities that are part of common ground but are not necessarily previously mentioned. They can, therefore, introduce new information.

(15) [Who is bothered by current problems?]
a. Насущные вопросы беспокоят [$_F$Левитана].
current questions.Nom worry Levitan.Acc
Lit. 'Levitan, pressing issues are bothering.'
b. ??Насущные вопросы беспокоят [$_F$его].
current questions.Nom worry him.Acc
Lit.: 'Him, pressing issues are bothering.'

This explains our second finding, namely, that pronominalization was a reliable indicator of subjecthood in SVO sentences and of objecthood in OVS sentences. Since pronouns convey given information, they are likely to be fronted regardless of their grammatical role. Proper nouns, on the other hand, tend to be subjects in both word order variations. Most of the time, we talk about people and their actions. Pronouns can refer to entities that have been already introduced in discourse, e.g. by a proper name. In SVO sentences subjects expressed by proper names can either be part of new information (or sentential focus when the whole sentence presents new information as an answer to the general question 'What happened?') or of the old information (that is, the topic). In OVS sentences, the subject is in focus; it must, therefore, convey new information. Moreover, foci in OVS sentences answer *wh*-questions that presuppose that, e.g., 'X did something' and X in such cases is likely to be an animate agent expressed by a proper name or a noun phrase, as in (16), but not a pronoun.

(16) a. Кощунственную погребальную затянул Исаак Ильич.
blasphemous dirge.Acc was.dragging Isaak Il'ič.Nom
'It was Isaak Il'ič who started a blasphemous dirge.' (Evdokimov 1959)
b. Кощунственную погребальную затянул какой-то мужик.
blasphemous dirge.Acc was.dragging some man.Nom
'It was some man who started a blasphemous dirge.'

Proper names are also lengthier than pronouns; besides actual names they often include titles. In general, we found a tendency for shorter constituents to follow longer constituents. This tendency can also be linked to the information structure of sentences. Since given information does not have to be described again, it can be expressed by short constituents. On the other hand, one might need longer constituents to introduce new information (cf. Arnold et al. 2000).

6 Conclusions

Previous studies suggested that such properties of constituents as animacy, definiteness and complexity (among others) can be used to distinguish between subjects and objects. Moreover, they can be used to predict grammatical functions of NPs. The results were based on the preferences found across languages for subjects to be more animate and definite than objects; and further on the assumption that non-typical constituents are unlikely to precede typical constituents to facilitate disambiguation of grammatical functions for the hearer. However, previous work analyzed languages with relatively fixed word orders. In such languages, the word order is primarily used to distinguish grammatical functions. Also the preference for animate constituents to precede inanimate ones coincides in such cases with subjects preceding non-subjects. But what happens when grammatical functions are indicated by case-marking and the word order is not fixed? This is the question we addressed in this study.

By comparing animacy, definiteness and the length of subjects and objects in a sample of Russian SVO and OVS sentences from two different genres, we were able to analyze these properties in relation to the grammatical functions, taking into account the word order of the sentences in which they occurred.

Our findings are similar to previous studies for the sets of SVO sentences only, in which subjects are more animate and definite than objects. However, in OVS sentences these properties were less reliable indicators of grammatical functions. Fronted objects were equally animate to subjects. Moreover, fronted objects were likely to be pronominalized—a property that is usually attributed to subjects. Going back to the question we raised at the beginning, namely, whether animacy and definiteness of constituents interplay with grammatical functions or information structure, our results suggest that the properties of constituents are linked to the information structure of sentences. Since pronouns are likely to refer to given information, they are likely to be fronted regardless of their grammatical functions. In languages with fixed word order, the initial position coincides with the subject position. When the word order is free, however, the initial position is not necessarily occupied by the subject. Rather, it is occupied by constituents that express given information. In OVS sentences such constituents are often expressed by means of pronouns. Our findings are relevant to existing accounts on constituent properties, their grammatical functions and the ordering in that they emphasize that word order variation can significantly influence the results.

Sources

Evdokimov, N. N. (1959). http://www.lib.ru/MEMUARY/ZHZL/lewitan.txt.
Penn Treebank corpus: www.ldc.upenn.edu/ldc/online/treebank

References

Aissen, J. (2003). Differential object marking: iconicity vs. economy. *Natural Language and Linguistic Theory*, *21*(3), 435–483.

Arnold, J. E. et al. (2000). Heaviness vs. newness: the effects of structural complexity and discourse status on constituent ordering. *Language*, *76*, 28–55.
Battistella, E. L. (1990). *Markedness. The evaluative superstructure of language*. Albany.
Bivon, R. (1971). *Element order* (Studies in the Modern Russian Language, *7*). New York.
Bouma, G. J. (2008). *Starting a sentence in Dutch: a corpus study of subject- and object-fronting*. Ph.D. dissertation. Groningen. http://dissertations.ub.rug.nl/faculties/arts/2008/g.j.bouma/. Accessed 21 Oct. 2010.
Clark, H. H., & Clark, E. V. (1977). *Psychology and language. An introduction to psycholinguistics*. New York.
Comrie, B. (1989). *Language universals and linguistic typology. Syntax and morphology.* Chicago.
Dahl, Ö. (2000). Egophoricity in discourse and syntax. *Functions of Language*, *7*(1), 37–77.
Dahl, Ö., & Fraurud, K. (1996). Animacy in grammar and discourse. In T. Fretheim & J. K. Gundel (Eds.), *Reference and referent accessibility* (Pragmatics & Beyond. New Series, *38*) (pp. 47–64). Amsterdam.
Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, *67*(3), 547–619.
Givón, T. (1983). Topic continuity in discourse: an introduction. In T. Givón (Ed.), *Topic continuity in discourse: a quantitative cross-language study* (Typological Studies in Language, *3*) (pp. 1–41). Amsterdam.
Givón, T. (2001). *Syntax: an introduction* (Vol. 1–2). Amsterdam.
Gundel, J. K. (1988). Universals of topic-comment structure. In M. Hammond, E. Moravcsik & J. Wirth (Eds.), *Studies in syntactic typology* (Typological Studies in Language, *17*) (pp. 209–239). Amsterdam.
Jacobsen, W. M. (1992). *The transitive structure of events in Japanese* (Studies in Japanese linguistics, *1*). Tokyo.
Jakobson, R. (1971[1936]). Beitrag zur allgemeinen Kasuslehre. Gesamtbedeutungen der russischen Kasus. In *Selected Writings*. *Volume II: Word and language* (pp. 23–71). Den Haag, Paris.
Hawkins, J. A. (1983). *Word order universals*. New York.
Heylen, K. (2005). A quantitative corpus study of German word order variation. In S. Kepser & M. Reis (Eds.), *Linguistic evidence. Empirical, theoretical and computational perspectives* (Studies in Generative Grammar, *85)* (pp. 241–263). Berlin.
King, T. H. (1995). *Configuring topic and focus in Russian*. Stanford.
Kovtunova, I. I. (1976). *Sovremennyj russkij jazyk. Porjadok slov i aktual'noe členenie predloženija*. Moskva.
Levin, B. (1993). *English verb classes and alternations: a preliminary investigation*. Chicago.
Marcus, M. P. et al. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.
Morimoto, Y. (2001). Verb raising and phrase structure variation in OT. In P. Sells (Ed.), *Formal and empirical issues in optimality theoretic syntax* (Studies in Constraint-Based Lexicalism, *5*) (pp. 129–196). Stanford.
Øvrelid, L. (2004). Disambiguation of syntactic functions in Norwegian: modeling variation in word order interpretations conditioned by animacy and definiteness. In F. Karlsson (Ed.), *Proceedings of the 20th Scandinavian conference of linguistics*. Helsinki. http://www.ling.helsinki.fi/kielitiede/20scl/Ovrelid.pdf. Accessed 21 October 2010.
Prince, A., & Smolensky, P. (1993). *Optimality theory: constraint interaction in generative grammar*. Piscataway.
Rosenbach, A. (2002). *Genitive variation in English. Conceptual factors in synchronic and diachronic studies* (Topics in English Linguistics, *42*). Berlin, New York.
Rosenbach, A. (2003). Aspects of iconicity and economy in the choice between the *s*-genitive and the *of*-genitive in English. In G. Rohdenburg & B. Mondorf (Eds.), *Determinants of grammatical variation in English* (Topics in English Linguistics, *43*) (pp. 379–411). Berlin, New York.
Siewierska, A. (1988). *Word order rules*. London.
Weber, A., & Müller, K. (2004). Word order variation in German main clauses: a corpus analysis. In S. Hansen-Schirra, S. Oepen & H. Uszkoreit (Eds.), *The 20th international conference on computational linguistics. Proceedings* (pp. 71–77). Geneva. http://acl.ldc.upenn.edu/coling2004/W8/pdf/proceedings.pdf. Accessed 21 October 2010.
Zeevat, H., & Jäger, G. (2002). A reinterpretation of syntactic alignment. In D. de Jongh, M. Nilsenová & H. Zeevat (Eds.), *Proceedings of the 3rd and 4th international symposium on language, logic and computation*. Amsterdam. http://www2.sfs.uni-tuebingen.de/jaeger/publications/jaeger_zeevat.pdf. Accessed 22 Nov. 2010.